(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2022년03월17일
(11) 등록번호 10-2375115
(24) 등록일자 2022년03월11일

(51) 국제특허분류(Int. Cl.)
G10L 15/06 (2006.01) *G10L 15/02* (2006.01)
G10L 15/16 (2006.01) *G10L 15/187* (2013.01)
G10L 15/193 (2013.01) *G10L 15/28* (2006.01)
G10L 15/30 (2013.01) *G10L 15/32* (2013.01)
(52) CPC특허분류
G10L 15/06 (2013.01)
G10L 15/02 (2013.01)
(21) 출원번호 10-2021-7035448
(22) 출원일자(국제) 2020년04월28일
심사청구일자 2021년10월29일
(85) 번역문제출일자 2021년10월29일
(65) 공개번호 10-2021-0138776
(43) 공개일자 2021년11월19일
(86) 국제출원번호 PCT/US2020/030321
(87) 국제공개번호 WO 2020/226948
국제공개일자 2020년11월12일
(30) 우선권주장
62/842,571 2019년05월03일 미국(US)
(56) 선행기술조사문헌
US20160104482 A1

(73) 특허권자
구글 엘엘씨
미국 캘리포니아 마운틴 뷰 엠피시어터 파크웨이 1600 (우:94043)
(72) 발명자
후, 커
미국 94043 캘리포니아주 마운틴 뷰 앰피시어터 파크웨이 1600
브루귀에르, 앙투안, 진
미국 94043 캘리포니아주 마운틴 뷰 앰피시어터 파크웨이 1600
(뒷면에 계속)
(74) 대리인
양영준, 이민호, 백만기

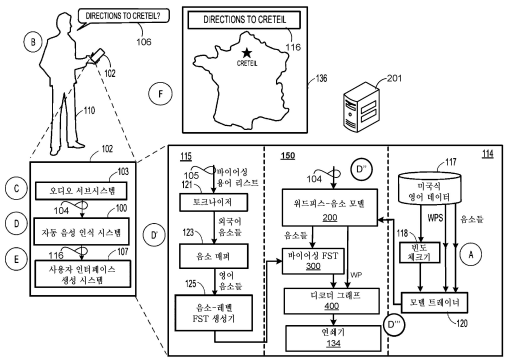
전체 청구항 수 : 총 20 항 심사관 : 강민정

(54) 발명의 명칭 **엔드-투-엔드 모델들에서 교차-언어 음성 인식을 위한 음소-기반 컨텍스트화**

(57) 요 약

방법(500)은 제1 언어의 원어민(110)에 의해 말해진 발화(106)를 인코딩하는 오디오 데이터를 수신하는 단계, 및 제1 언어와 상이한 제2 언어로 된 하나 이상의 용어를 포함하는 바이어싱 용어 리스트(105)를 수신하는 단계를 포함한다. 방법은 또한, 음성 인식 모델(200)을 사용하여, 제1 언어의 워드피스들 및 대응하는 음소 시퀀스들 둘 다에 대한 음성 인식 스코어들을 생성하기 위해 오디오 데이터로부터 도출된 음향 피처들(105)을 프로세싱하는 단계를 포함한다. 방법은 또한, 바이어싱 용어 리스트의 하나 이상의 용어에 기초하여 음소 시퀀스들에 대한 음성 인식 스코어들을 재스코어링하는 단계, 및 워드피스들에 대한 음성 인식 스코어들 및 음소 시퀀스들에 대한 재스코어링된 음성 인식 스코어들을 사용하여, 발화에 대한 전사(116)를 생성하기 위해 디코딩 그래프(400)를 실행하는 단계를 포함한다.

대 표 도



(52) CPC특허분류
G10L 15/063 (2013.01)
G10L 15/16 (2013.01)
G10L 15/187 (2013.01)
G10L 15/193 (2013.01)
G10L 15/285 (2013.01)
G10L 15/30 (2013.01)
G10L 15/32 (2013.01)
G10L 2015/025 (2013.01)
(72) 발명자
사이너스, 타라, 엔.
미국 94043 캘리포니아주 마운틴 뷰 앰피시어터 파크웨이 1600
프라브하발카르, 로히트, 프라카쉬
미국 94043 캘리포니아주 마운틴 뷰 앰피시어터 파크웨이 1600
푼닥, 고란
미국 94043 캘리포니아주 마운틴 뷰 앰피시어터 파크웨이 1600

명 세 서

청구범위

청구항 1

방법(500)으로서,

데이터 프로세싱 하드웨어(610)에서, 제1 언어의 원어민(110)에 의해 말해진 발화(utterance)(106)를 인코딩하는 오디오 데이터를 수신하는 단계;

상기 데이터 프로세싱 하드웨어(610)에서, 상기 제1 언어와 상이한 제2 언어로 된 하나 이상의 용어를 포함하는 바이어싱 용어 리스트(biasing term list)(105)를 수신하는 단계;

상기 데이터 프로세싱 하드웨어(610)에 의해, 음성 인식 모델(200)을 사용하여, 상기 제1 언어의 워드피스(wordpiece)들 및 대응하는 음소 시퀀스(phoneme sequence)들 둘 다에 대한 음성 인식 스코어들을 생성하기 위해 상기 오디오 데이터로부터 도출된 음향 피처(acoustic feature)들(104)을 프로세싱하는 단계;

상기 데이터 프로세싱 하드웨어(610)에 의해, 상기 바이어싱 용어 리스트(105)의 하나 이상의 용어에 기초하여 상기 음소 시퀀스들에 대한 음성 인식 스코어들을 재스코어링하는 단계; 및

상기 데이터 프로세싱 하드웨어(610)에 의해, 상기 워드피스들에 대한 음성 인식 스코어들 및 상기 음소 시퀀스들에 대한 재스코어링된 음성 인식 스코어들을 사용하여, 상기 발화(106)에 대한 전사(transcription)(116)를 생성하기 위해 디코딩 그래프(400)를 실행하는 단계

를 포함하는, 방법(500).

청구항 2

제1항에 있어서, 상기 음소 시퀀스들에 대한 음성 인식 스코어들을 재스코어링하는 단계는 바이어싱 유한-상태 변환기(finite-state transducer)(FST)를 사용하여 상기 음소 시퀀스들에 대한 음성 인식 스코어들을 재스코어링하는 단계를 포함하는, 방법(500).

청구항 3

제2항에 있어서,

상기 데이터 프로세싱 하드웨어(610)에 의해, 상기 바이어싱 용어 리스트(105)의 각각의 용어를 상기 제2 언어의 대응하는 음소 시퀀스로 토큰화하는 단계;

상기 데이터 프로세싱 하드웨어(610)에 의해, 상기 제2 언어의 각각의 대응하는 음소 시퀀스를 상기 제1 언어의 대응하는 음소 시퀀스에 매핑하는 단계; 및

상기 데이터 프로세싱 하드웨어(610)에 의해, 상기 제1 언어의 각각의 대응하는 음소 시퀀스에 기초하여 상기 바이어싱 FST를 생성하는 단계

를 추가로 포함하는, 방법(500).

청구항 4

제1항 내지 제3항 중 어느 한 항에 있어서, 상기 음성 인식 모델(200)은 엔드-투-엔드 워드피스-음소 모델(end-to-end, wordpiece-phoneme model)(200)을 포함하는, 방법(500).

청구항 5

제4항에 있어서, 상기 엔드-투-엔드 워드피스-음소 모델(200)은 순환 신경망-변환기(recurrent neural network-transducer)(RNN-T)를 포함하는, 방법(500).

청구항 6

제1항 내지 제3항 중 어느 한 항에 있어서, 상기 디코딩 그래프(400)의 실행 동안, 상기 디코딩 그래프(400)는 상기 바이어싱 용어 리스트(105)의 하나 이상의 용어 중 임의의 것을 선호(favor)하도록 상기 전사(116)를 바이어싱하는, 방법(500).

청구항 7

제1항 내지 제3항 중 어느 한 항에 있어서, 상기 음성 인식 모델(200)은 상기 제1 언어의 트레이닝 발화들에 대해서만 트레이닝되는, 방법(500).

청구항 8

제1항 내지 제3항 중 어느 한 항에 있어서, 상기 바이어싱 용어 리스트(105)의 용어들 중 어느 것도 상기 음성 인식 모델(200)을 트레이닝하는 데 사용되지 않은, 방법(500).

청구항 9

제1항 내지 제3항 중 어느 한 항에 있어서, 상기 데이터 프로세싱 하드웨어(610) 및 상기 음성 인식 모델(200)은 사용자 디바이스(102) 상에 상주하는, 방법(500).

청구항 10

제1항 내지 제3항 중 어느 한 항에 있어서,

상기 데이터 프로세싱 하드웨어(610) 및 상기 음성 인식 모델(200)은 원격 컴퓨팅 디바이스(201) 상에 상주하고,

상기 발화(106)를 인코딩하는 오디오 데이터를 수신하는 단계는 상기 원격 컴퓨팅 디바이스(201)와 통신하는 사용자 디바이스(102)로부터 상기 발화(106)를 인코딩하는 오디오 데이터를 수신하는 단계를 포함하는, 방법(500).

청구항 11

시스템(100)으로서,

데이터 프로세싱 하드웨어(610); 및

상기 데이터 프로세싱 하드웨어(610)와 통신하는 메모리 하드웨어(620) - 상기 메모리 하드웨어는, 상기 데이터 프로세싱 하드웨어(610) 상에서 실행될 때, 상기 데이터 프로세싱 하드웨어(610)로 하여금 동작들을 수행하게 하는 명령어들을 저장함 -

를 포함하고,

상기 동작들은,

제1 언어의 원어민(110)에 의해 말해진 발화(106)를 인코딩하는 오디오 데이터를 수신하는 동작;

상기 제1 언어와 상이한 제2 언어로 된 하나 이상의 용어를 포함하는 바이어싱 용어 리스트(105)를 수신하는 동작;

음성 인식 모델(200)을 사용하여, 상기 제1 언어의 워드피스들 및 대응하는 음소 시퀀스들 둘 다에 대한 음성 인식 스코어들을 생성하기 위해 상기 오디오 데이터로부터 도출된 음향 피처들(104)을 프로세싱하는 동작;

상기 바이어싱 용어 리스트(105)의 하나 이상의 용어에 기초하여 상기 음소 시퀀스들에 대한 음성 인식 스코어들을 재스코어링하는 동작; 및

상기 워드피스들에 대한 음성 인식 스코어들 및 상기 음소 시퀀스들에 대한 재스코어링된 음성 인식 스코어들을 사용하여, 상기 발화(106)에 대한 전사(116)를 생성하기 위해 디코딩 그래프(400)를 실행하는 동작

을 포함하는, 시스템(100).

청구항 12

제11항에 있어서, 상기 음소 시퀀스들에 대한 음성 인식 스코어들을 재스코어링하는 동작은 바이어싱 유한-상태

변환기(FST)를 사용하여 상기 음소 시퀀스들에 대한 음성 인식 스코어들을 재스코어링하는 동작을 포함하는, 시스템(100).

청구항 13

제12항에 있어서, 상기 동작들은,

상기 바이어싱 용어 리스트(105)의 각각의 용어를 상기 제2 언어의 대응하는 음소 시퀀스로 토큰화하는 동작;

상기 제2 언어의 각각의 대응하는 음소 시퀀스를 상기 제1 언어의 대응하는 음소 시퀀스에 매핑하는 동작; 및

상기 제1 언어의 각각의 대응하는 음소 시퀀스에 기초하여 상기 바이어싱 FST를 생성하는 동작

을 추가로 포함하는, 시스템(100).

청구항 14

제11항 내지 제13항 중 어느 한 항에 있어서, 상기 음성 인식 모델(200)은 엔드-투-엔드 워드피스-음소 모델(200)을 포함하는, 시스템(100).

청구항 15

제14항에 있어서, 상기 엔드-투-엔드 워드피스-음소 모델(200)은 순환 신경망-변환기(RNN-T)를 포함하는, 시스템(100).

청구항 16

제11항 내지 제13항 중 어느 한 항에 있어서, 상기 디코딩 그래프(400)의 실행 동안, 상기 디코딩 그래프(400)는 상기 바이어싱 용어 리스트(105)의 하나 이상의 용어 중 임의의 것을 선호하도록 상기 전사(116)를 바이어싱하는, 시스템(100).

청구항 17

제11항 내지 제13항 중 어느 한 항에 있어서, 상기 음성 인식 모델(200)은 상기 제1 언어의 트레이닝 발화들에 대해서만 트레이닝되는, 시스템(100).

청구항 18

제11항 내지 제13항 중 어느 한 항에 있어서, 상기 바이어싱 용어 리스트(105)의 용어들 중 어느 것도 상기 음성 인식 모델(200)을 트레이닝하는 데 사용되지 않은, 시스템(100).

청구항 19

제11항 내지 제13항 중 어느 한 항에 있어서, 상기 데이터 프로세싱 하드웨어(610) 및 상기 음성 인식 모델(200)은 사용자 디바이스(102) 상에 상주하는, 시스템(100).

청구항 20

제11항 내지 제13항 중 어느 한 항에 있어서,

상기 데이터 프로세싱 하드웨어(610) 및 상기 음성 인식 모델(200)은 원격 컴퓨팅 디바이스(201) 상에 상주하고,

상기 발화(106)를 인코딩하는 오디오 데이터를 수신하는 동작은 상기 원격 컴퓨팅 디바이스(201)와 통신하는 사용자 디바이스(102)로부터 상기 발화(106)를 인코딩하는 오디오 데이터를 수신하는 동작을 포함하는, 시스템(100).

발명의 설명

기 술 분 야

[0001] 본 개시내용은 엔드-투-엔드(end-to-end) 모델들에서 교차-언어 음성 인식(cross-lingual speech recognitio

n)을 위한 음소-기반 컨텍스트화(phoneme-based contextualization)에 관한 것이다.

배 경 기 술

[0002] 음성 컨텍스트를 인식하는 것은 자동 음성 인식(automatic speech recognition)(ASR) 시스템들의 목표이다. 그러나, 사람들이 말할 수 있는 워드들이 매우 다양하고 억양들 및 발음의 많은 변형들이 있다는 것을 고려하면 음성의 컨텍스트를 인식하는 능력은 어려운 일이다. 많은 경우들에서, 사람이 말하는 워드들 및 구문들의 타입들은 사람이 처한 컨텍스트에 따라 달라진다.

[0003] 컨텍스트 자동 음성 인식(ASR)은 사용자 자신의 재생 리스트, 연락처들 또는 지리적 장소 이름들과 같은 주어진 컨텍스트에 대한 음성 인식을 바이어싱(biasing)시키는 것을 포함한다. 컨텍스트 정보는 대개 인식해야 하는 관련 구문들의 리스트를 포함하며, 이는 종종 트레이닝에서 드물게 보이는 희귀 구문들 또는 외국어 워드들도 포함한다. 컨텍스트 바이어싱을 수행하기 위해, 종래의 ASR 시스템들은 때때로 n-gram 가중 유한 상태 변환기(weighted finite state transducer)(WFST)를 사용하여 독립적인 컨텍스트 언어 모델(language model)(LM)에서 컨텍스트 정보를 모델링하고, 온-더-플라이(on-the-fly)(OTF) 리스코어링에 대한 기준선 LM으로 독립적인 컨텍스트 LM을 구성한다.

[0004] 최근에, 엔드-투-엔드(end-to-end)(E2E) 모델들은 ASR에 대한 큰 가능성을 보여주었으며, 종래의 온-디바이스 모델들과 비교하여 향상된 워드 오류율(word error rate)들(WER들) 및 레이턴시 메트릭들을 나타낸다. 음성-대-텍스트 매핑을 직접 학습하기 위해 음향 모델(acoustic model)(AM), 발음 모델(pronunciation model)(PM) 및 LM들을 단일 네트워크로 폴딩하는 이러한 E2E 모델들은 별도의 AM, PM 및 LM들을 갖는 종래의 ASR 시스템들과 비교하여 경쟁력 있는 결과들을 보여주었다. 대표적인 E2E 모델들은 워드-기반 연결성 시계열 분류(connectionist temporal classification)(CTC) 모델들, 순환 신경망 변환기(recurrent neural network transducer)(RNN-T) 모델들, 및 듣기, 집중하기 및 철자 맞추기(Listen, Attend, and Spell)(LAS)와 같은 주의-기반 모델들을 포함한다. E2E 모델들은 빔-검색 디코딩 동안 제한된 수의 인식 후보들을 유지하기 때문에, 컨텍스트 ASR은 E2E 모델들에서 어려울 수 있다.

발명의 내용

[0005] 본 개시내용의 일 양태는 바이어싱 용어 리스트에 존재하는 용어들에 대한 음성 인식 결과들을 바이어싱하기 위한 방법을 제공한다. 방법은, 데이터 프로세싱 하드웨어에서, 제1 언어의 원어민에 의해 말해진 발화(utterance)를 인코딩하는 오디오 데이터를 수신하는 단계, 및 데이터 프로세싱 하드웨어에서, 제1 언어와 상이한 제2 언어로 된 하나 이상의 용어를 포함하는 바이어싱 용어 리스트(biasing term list)를 수신하는 단계를 포함한다. 방법은 또한, 데이터 프로세싱 하드웨어에서, 음성 인식 모델을 사용하여, 제1 언어의 워드피스(wordpiece)들 및 대응하는 음소 시퀀스(phoneme sequence)들 둘 다에 대한 음성 인식 스코어들을 생성하기 위해 오디오 데이터로부터 도출된 음향 피처(acoustic feature)들을 프로세싱하는 단계를 포함한다. 방법은 또한, 데이터 프로세싱 하드웨어에 의해, 바이어싱 용어 리스트의 하나 이상의 용어에 기초하여 음소 시퀀스들에 대한 음성 인식 스코어들을 재스코어링하는 단계를 포함한다. 방법은 또한, 데이터 프로세싱 하드웨어에 의해, 워드피스들에 대한 음성 인식 스코어들 및 음소 시퀀스들에 대한 재스코어링된 음성 인식 스코어들을 사용하여, 발화에 대한 전사(transcription)를 생성하기 위해 디코딩 그래프를 실행하는 단계를 포함한다.

[0006] 본 개시내용의 구현들은 다음의 임의적인 피처들 중 하나 이상을 포함할 수 있다. 일부 구현들에서, 음소 시퀀스들에 대한 음성 인식 스코어들을 재스코어링하는 단계는 바이어싱 유한-상태 변환기(finite-state transducer)(FST)를 사용하여 음소 시퀀스들에 대한 음성 인식 스코어들을 재스코어링하는 단계를 포함한다. 이러한 구현들에서, 방법은 또한, 데이터 프로세싱 하드웨어에 의해, 바이어싱 용어 리스트의 각각의 용어를 제2 언어의 대응하는 음소 시퀀스로 토큰화하는 단계; 데이터 프로세싱 하드웨어에 의해, 제2 언어의 각각의 대응하는 음소 시퀀스를 제1 언어의 대응하는 음소 시퀀스에 매핑하는 단계; 및 데이터 프로세싱 하드웨어에 의해, 제1 언어의 각각의 대응하는 음소 시퀀스에 기초하여 바이어싱 FST를 생성하는 단계를 포함할 수 있다.

[0007] 일부 예들에서, 음성 인식 모델은 엔드-투-엔드 워드피스-음소 모델(end-to-end, wordpiece-phoneme model)을 포함한다. 특정 예에서, 엔드-투-엔드 워드피스-음소 모델은 순환 신경망-변환기(recurrent neural network-transducer)(RNN-T)를 포함한다.

[0008] 일부 구현들에서, 디코딩 그래프의 실행 동안, 디코딩 그래프는 바이어싱 용어 리스트의 하나 이상의 용어 중 임의의 것을 선호(favor)하도록 전사를 바이어싱한다. 음성 인식 모델은 제1 언어의 트레이닝 발화들에 대해서

만 트레이닝될 수 있다. 또한, 바이어싱 용어 리스트의 용어들 중 어느 것도 음성 인식 모델을 트레이닝하는 데 사용되지 않을 수 있다.

[0009] 데이터 프로세싱 하드웨어 및 음성 인식 모델은 사용자 디바이스 또는 사용자 디바이스와 통신하는 원격 컴퓨팅 디바이스 상에 상주할 수 있다. 데이터 프로세싱 하드웨어 및 음성 인식 모델이 원격 컴퓨팅 디바이스 상에 상주할 때, 발화를 인코딩하는 오디오 데이터를 수신하는 단계는 사용자 디바이스로부터 발화를 인코딩하는 오디오 데이터를 수신하는 단계를 포함할 수 있다.

[0010] 본 개시내용의 다른 양태는 바이어싱 용어 리스트에 존재하는 용어들에 대한 음성 인식 결과들을 바이어싱하기 위한 시스템을 제공한다. 시스템은 데이터 프로세싱 하드웨어, 및 데이터 프로세싱 하드웨어와 통신하고, 데이터 프로세싱 하드웨어 상에서 실행될 때, 데이터 프로세싱 하드웨어로 하여금 동작들을 수행하게 하는 명령어들을 저장하는 메모리 하드웨어를 포함한다. 동작들은 제1 언어의 원어민에 의해 말해진 발화를 인코딩하는 오디오 데이터를 수신하는 동작; 제1 언어와 상이한 제2 언어로 된 하나 이상의 용어를 포함하는 바이어싱 용어 리스트를 수신하는 동작; 및 음성 인식 모델을 사용하여, 제1 언어의 워드피스들 및 대응하는 음소 시퀀스들 둘 다에 대한 음성 인식 스코어들을 생성하기 위해 오디오 데이터로부터 도출된 음향 피처들을 프로세싱하는 동작을 포함한다. 동작들은 또한 바이어싱 용어 리스트의 하나 이상의 용어에 기초하여 음소 시퀀스들에 대한 음성 인식 스코어들을 재스코어링하는 동작, 및 워드피스들에 대한 음성 인식 스코어들 및 음소 시퀀스들에 대한 재스코어링된 음성 인식 스코어들을 사용하여, 발화에 대한 전사를 생성하기 위해 디코딩 그래프를 실행하는 동작을 포함한다.

[0011] 이 양태는 다음의 임의적인 피처들 중 하나 이상을 포함할 수 있다. 일부 구현들에서, 음소 시퀀스들에 대한 음성 인식 스코어들을 재스코어링하는 동작은 바이어싱 유한-상태 변환기(FST)를 사용하여 음소 시퀀스들에 대한 음성 인식 스코어들을 재스코어링하는 동작을 포함한다. 이러한 구현들에서, 동작들은 또한 바이어싱 용어 리스트의 각각의 용어를 제2 언어의 대응하는 음소 시퀀스로 토큰화하는 동작; 제2 언어의 각각의 대응하는 음소 시퀀스를 제1 언어의 대응하는 음소 시퀀스에 매핑하는 동작; 및 제1 언어의 각각의 대응하는 음소 시퀀스에 기초하여 바이어싱 FST를 생성하는 동작을 포함할 수 있다.

[0012] 일부 예들에서, 음성 인식 모델은 엔드-투-엔드 워드피스-음소 모델을 포함한다. 특정 예에서, 엔드-투-엔드 워드피스-음소 모델은 순환 신경망-변환기(RNN-T)를 포함한다.

[0013] 일부 구현들에서, 디코딩 그래프의 실행 동안, 디코딩 그래프는 바이어싱 용어 리스트의 하나 이상의 용어 중 임의의 것을 선호하도록 전사를 바이어싱한다. 음성 인식 모델은 제1 언어의 트레이닝 발화들에 대해서만 트레이닝될 수 있다. 또한, 바이어싱 용어 리스트의 용어들 중 어느 것도 음성 인식 모델을 트레이닝하는 데 사용되지 않을 수 있다.

[0014] 데이터 프로세싱 하드웨어 및 음성 인식 모델은 사용자 디바이스 또는 사용자 디바이스와 통신하는 원격 컴퓨팅 디바이스 상에 상주할 수 있다. 데이터 프로세싱 하드웨어 및 음성 인식 모델이 원격 컴퓨팅 디바이스 상에 상주할 때, 발화를 인코딩하는 오디오 데이터를 수신하는 동작은 사용자 디바이스로부터 발화를 인코딩하는 오디오 데이터를 수신하는 동작을 포함할 수 있다.

[0015] 본 개시내용의 하나 이상의 구현의 세부사항은 첨부 도면들 및 하기 설명에 개시되어 있다. 다른 양태들, 피처들 및 이점들은 설명 및 도면들, 및 청구범위로부터 명백할 것이다.

도면의 간단한 설명

[0016] 도 1은 바이어싱 용어 리스트에 존재하는 용어들에 대한 음성 인식 결과들을 바이어싱하는 음성 인식 모델을 포함하는 예시적인 자동 음성 인식 시스템의 개략도이다.

도 2는 도 1의 음성 인식 모델의 예시적인 아키텍처의 개략도이다.

도 3은 예시적인 바이어싱 유한-상태 변환기의 개략도이다.

도 4는 워드피스들 및 대응하는 음소 시퀀스들에 기초한 예시적인 디코딩 그래프의 개략도이다.

도 5는 바이어싱 용어 리스트에 존재하는 용어들에 대한 음성 인식 결과들을 바이어싱하는 방법에 대한 예시적인 동작들의 배열의 흐름도이다.

도 6은 본 명세서에 설명된 시스템들 및 방법들을 구현하는 데 사용될 수 있는 예시적인 컴퓨팅 디바이스의 개

략도이다.

다양한 도면들에서 유사한 참조 심볼들은 유사한 요소들을 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0017] 본 명세서의 구현들은 컨텍스트 자동 음성 인식(ASR) 모델을 강화시켜, 다른 동작들 중에서, 외국어 음소 세트를 ASR 모델의 언어(예를 들어, 미국식 영어)에 대한 음소 세트에 매핑함으로써 외국어 워드들을 인식하여, 음소-레벨 바이어싱 유한 상태 변환기(FST)에서 외국어 워드들의 모델링을 가능하게 하는 것에 관한 것이다. 추가 구현들은 모델링 공간에 ASR 모델의 언어(예를 들어, 미국식 영어)에 대한 워드피스들 및 음소들을 포함하는 워드피스-음소 모델을 통합하는 ASR 모델에 관한 것이다. 예를 들어, 컨텍스트 ASR 모델은 워드피스-음소 모델 및 컨텍스트 바이어싱 FST를 사용하여 음성 발화를 디코딩하여 하나 이상의 외국어 워드에 대한 발화의 전사를 컨텍스트적으로 바이어싱하도록 구성된다. 예를 들어, 미국식 영어 화자는 Crétteil이라는 워드가 프랑스어 워드인 "Directions to Crétteil"라는 발화를 말할 수 있고, 워드피스 음소 모델 및 컨텍스트 바이어싱 FST를 활용하는 컨텍스트 ASR 모델은 컨텍스트 ASR 모델이 미국식 영어 이외의 언어들로 된 임의의 워드들에 대해 트레이닝된 적이 없음에도 불구하고 Crétteil이라는 외국어 워드를 인식하도록 전사를 바이어싱할 수 있다. 이 예에서, Crétteil이라는 외국어 워드는 현재 컨텍스트에 기초한 바이어싱 워드 리스트에 포함된 다수의 프랑스어 워드들 중 하나일 수 있다. 예를 들어, 사용자가 현재 프랑스에 있고 운전 중인 경우, 현재 컨텍스트는 프랑스 도시/지역 이름들이 관련이 있음을 나타낼 수 있고, 따라서 컨텍스트 ASR 모델은 이러한 프랑스 도시/지역 이름들에 대해 바이어싱될 수 있다.

[0018] 도 1을 참조하면, 일부 구현들에서, 강화된 ASR 시스템(100)은 외국어 워드들을 인식하도록 강화된다. 도시된 예에서, ASR 시스템(100)은 사용자(110)의 사용자 디바이스(102) 및/또는 사용자 디바이스와 통신하는 원격 컴퓨팅 디바이스(201)(예를 들어, 클라우드-컴퓨팅 환경에서 실행되는 분산 시스템의 하나 이상의 서버(serve)) 상에 상주한다. 사용자 디바이스(102)가 모바일 컴퓨팅 디바이스(예를 들어, 스마트폰)로서 도시되어 있지만, 사용자 디바이스(102)는 제한 없이 태블릿 디바이스, 랩탑/데스크탑 컴퓨터, 웨어러블 디바이스, 디지털 어시스턴트 디바이스, 스마트 스피커/디스플레이, 스마트 기기, 자동차 인포테인먼트 시스템 또는 사물 인터넷(Internet-of-Things)(IoT) 디바이스와 같은 임의의 타입의 컴퓨팅 디바이스에 대응할 수 있다.

[0019] 사용자 디바이스(102)는 사용자(104)에 의해 말해진 발화(106)를 수신하고(예를 들어, 사용자 디바이스(102)는 말해진 발화(106)를 녹음하기 위한 하나 이상의 마이크로폰을 포함할 수 있음), 발화(106)를 ASR 시스템(100)에 의해 프로세싱될 수 있는 파라미터화된 입력 음향 프레임들(104)과 연관된 대응하는 디지털 포맷으로 변환하도록 구성되는 오디오 서브시스템(103)을 포함한다. 도시된 예에서, 사용자는 "Directions to Crétteil"이라는 구문에 대한 개개의 발화(106)를 말하고, 오디오 서브시스템(108)은 발화(106)를 ASR 시스템(100)에 대한 입력을 위해 대응하는 음향 프레임들(104)로 변환한다. 예를 들어, 음향 프레임들(104)은 80-차원 log-Mel 피처들을 각각 포함하는 일련의 파라미터화된 입력 음향 프레임들일 수 있으며, 이들은 짧은, 예를 들어, 25ms의 윈도우로 계산되고, 몇 밀리초마다, 예를 들어, 10밀리초마다 시프트될 수 있다.

[0020] 그 후, ASR 시스템(100)은 발화(106)에 대응하는 음향 프레임들(104)을 입력으로서 수신하고, 발화(106)에 대한 대응하는 전사(예를 들어, 인식 결과/가설)(116)를 출력으로서 생성/예측한다. 도시된 예에서, 사용자 디바이스(102) 및/또는 원격 컴퓨팅 디바이스(201)는 또한 사용자 디바이스(102)의 사용자 인터페이스(136)에서 사용자(104)에게 발화(106)의 전사(116)의 표현을 제시하도록 구성되는 사용자 인터페이스 생성기(107)를 실행한다. 일부 예들에서, 사용자 인터페이스(136)는 사용자 디바이스(102)와 통신하는 스크린 상에 디스플레이될 수 있다.

[0021] 일부 구성들에서, ASR 시스템(100)으로부터 출력된 전사(116)는, 예를 들어, 사용자 커맨드를 실행하기 위해 사용자 디바이스(102) 또는 원격 컴퓨팅 디바이스(201) 상에서 실행되는 자연어 이해(natural language understanding)(NLU) 모듈에 의해 프로세싱된다. 추가적으로 또는 대안적으로, 텍스트-대-음성 시스템(text-to-speech system)(예를 들어, 사용자 디바이스(102) 또는 원격 컴퓨팅 디바이스(201)의 임의의 조합에서 실행됨)은 다른 디바이스에 의한 가청 출력을 위해 전사를 합성된 음성으로 변환할 수 있다. 예를 들어, 원래 발화(106)는 사용자(104)가 친구에게 전송하고 있는 메시지에 대응할 수 있으며, 원래 발화(106)에서 전달되는 메시

지를 듣기 위한 친구에 대한 가청 출력을 위해 전사(116)가 합성된 음성으로 변환된다.

[0022] 강화된 ASR 시스템(100)은 바이어싱 컴포넌트(115), 워드피스-음소 모델(200) 및 바이어싱 FST(300)를 갖는 음성 인식기(150), 및 트레이닝 컴포넌트(114)를 포함한다. 바이어싱 컴포넌트(115)는 바이어싱 FST(300)를 생성하도록 구성되고, 트레이닝 컴포넌트(114)는 음소 레벨에서 외국어 워드들을 재스코어링함으로써 컨텍스트 바이어싱을 수행하도록 워드피스-음소 모델(200) 및 바이어싱 FST(300)를 트레이닝하도록 구성된다. 명백해지는 바와 같이, 음성 인식기(150)는 트레이닝된 워드피스-음소 모델(200) 및 바이어싱 FST(300)를 사용하여 외국어 워드들에 대해 바이어싱함으로써 컨텍스트 음성 인식을 수행한다.

[0023] 트레이닝 컴포넌트(114)는 단일 언어, 예를 들어, 미국식 영어로 된 텍스트의 코퍼스(corpus)를 갖는 어휘집(117), 빈도 체크기(118), 및 모델 트레이너(120)를 포함한다. 빈도 체크기(118)는 코퍼스의 텍스트 중 단일 언어로 된 용어들의 상대적 빈도를 결정하도록 구성되고, 모델 트레이너(120)는 텍스트 코퍼스의 용어들의 워드피스들 및 음소들 둘 다에 기초하여 워드피스-음소 모델(200)을 트레이닝하여 모델링 공간에 워드피스들 및 음소들 둘 다를 포함하도록 구성된다. 일부 예들에서, 워드피스-음소 모델(200)은 단일 언어, 예를 들어, 미국식 영어로부터의 워드피스-음소 세트들만을 포함하고 다른 언어들로부터의 워드피스-음소 세트들은 제외하는 트레이닝 데이터를 사용하여 모델 트레이너(120)에 의해 엔드-투-엔드 방식으로 트레이닝된다. 모델 트레이너(120)는 워드-빈도 기반 샘플링 전략을 채택하여 어휘집(117)을 사용하여 타겟 시퀀스에서 희귀 워드들을 음소들로 랜덤하게 토큰화할 수 있다. 스테이지 A에서, 트레이닝 컴포넌트(114)는 어휘집(117)으로부터의 텍스트를 사용하여 워드피스-음소 모델(200)을 트레이닝한다.

[0024] 일부 예들에서, 어휘집(117)은 약 50만 워드를 포함하며, 이들의 빈도들은 음소 시퀀스들을 사용할 때를 결정하는 데 사용된다. 어휘집(117)은 트레이닝 데이터로부터의 워드들 및 그들의 빈도들을 포함하고, 동음이의어들(예를 들어, "flower" 및 "flour"), 동형이의어들(예를 들어, 동사 또는 형용사로서의 "live") 및 발음 변형들(예를 들어, "either")을 제거함으로써 트리밍된다. 따라서, 어휘집(117)은 철자로부터 발음으로 또는 그 반대로 이동할 때 모호하지 않은 항목들만을 포함한다.

[0025] 일부 구현들에서, 모델 트레이너(120)는 트레이닝 입력 발화들을 25ms 프레임들로 분할하며, 이들은 10ms의 레이트로 윈도우화되고 시프트된다. 각각의 프레임에서 80-차원 log-Mel 피처가 추출되고, 현재 프레임과 왼쪽의 두 프레임이 연쇄(concatenate)되어 240-차원 log-Mel 피처를 생성한다. 그런 다음, 이러한 피처들은 30ms의 레이트로 다운-샘플링된다.

[0026] 일부 구현들에서, 워드피스-음소 모델(200)은 시퀀스-대-시퀀스 모델(sequence-to-sequence model)을 포함한다. 일부 예들에서, 워드피스-음소 모델(200)은 순환 신경망-변환기(RNN-T) 시퀀스-대-시퀀스 모델 아키텍처를 포함한다. 다른 예들에서, 모델(200)은 모델 아키텍처를 시퀀스화하기 위해 듣기, 집중하기, 철자 맞추기(Listen, Attend, Spell) 시퀀스를 포함한다.

[0027] 워드피스-음소 모델(200)은 트레이닝 시 소수의 워드들이 선택적으로 음소들로 분해될 수 있다는 점에서 워드피스-전용 모델과 상이하다. 모델의 출력은 심볼 세트가 워드피스와 음소 심볼들의 조합(union)인 단일 소프트맥스이다. 발음 어휘집이 워드들의 음소 시퀀스들을 획득하는 데 사용된다. 음소들은 희귀 워드들을 인식하는 데 강점을 보이기 때문에, 이러한 워드들은 더 자주 음소들로서 제시된다. 타겟 문장에서, i번째 워드는 확률 $p(i) = p_0 \cdot \min(\frac{T}{c(i)}, 1.0)$ 인 음소들로서 랜덤하게 제시되며, 여기서 p_0와 T는 상수들이고, c(i)는 전체 트레이닝 코퍼스에 워드가 나타나는 횟수를 표현하는 정수이다. T회 이하로 나타나는 워드들은 확률 p_0인 음소들로서 제시될 것이다. T회보다 많이 나타나는 워드들의 경우, 빈도가 높을수록 음소들로서 제시될 가능성이 적다. 일부 예들에서, T는 10과 같고 p_0는 5와 같지만, 다른 예들에서는, 상이한 값들이 선택될 수 있다. 워드피스들을 사용할지 음소들을 사용할지 결정하는 것은 각각의 기울기 반복(gradient iteration)에서 랜덤하게 이루어지고, 따라서 주어진 문장은 상이한 시기들에서 상이한 타겟 시퀀스들을 가질 수 있다는 것에 유의하도록 한다. 일부 구현들에서, 음소들은 컨텍스트-독립적 음소들이다.

[0028] 도 2를 참조하면, 워드피스-음소 모델(200)은 상호 작용형 애플리케이션(interactive application)들과 연관된 레이턴시 제한들을 준수하는 엔드-투-엔드(E2E) RNN-T 모델(200)을 포함할 수 있다. RNN-T 모델(200)은 작은 계산 풋프린트를 제공하고 종래의 ASR 아키텍처들보다 적은 메모리 요구사항들을 활용하여, RNN-T 모델 아키텍처를 사용자 디바이스(102) 상에서 완전히 음성 인식을 수행하는 데 적절하게 만든다(예를 들어, 원격 서버와의 통신이 필요하지 않다). RNN-T 모델(200)은 인코더 네트워크(210), 예측 네트워크(220) 및 조인트 네트워크

(230)를 포함한다. 전통적인 ASR 시스템의 음향 모델(acoustic model)(AM)과 대략 유사한 인코더 네트워크(210)는 스택형 장단기 메모리(Long Short-Term Memory)(LSTM) 계층들의 순환 네트워크를 포함한다. 예를 들어, 인코더는 d-차원 피처 벡터들(예를 들어, 음향 프레임들(104)(도 1))의 시퀀스 $x = (x_1, x_2, \cdots, x_T)$를 판독하고 - $x_t \in \mathbb{R}^d$ 임 -, 각각의 시간 단계에서 고차 피처 표현(higher-order feature representation)을 생성한다. 이 고차 피처 표현은 $\mathbf{h}_1^{\text{enc}}, \cdots, \mathbf{h}_T^{\text{enc}}$ 로서 표시된다.

[0029] 유사하게, 예측 네트워크(220)는 또한, 언어 모델(LM)과 같이, 지금까지 최종 소프트맥스(Softmax) 계층(240)에 의해 출력된 비-공백 심볼들의 시퀀스 $y_0, \ldots, y_{ui-1}$를 밀집 표현(dense representation) $\mathbf{p}_{u_i}$로 프로세싱하는 LSTM 네트워크이다. 마지막으로, RNN-T 모델 아키텍처를 사용하여, 인코더 및 예측 네트워크들(210, 220)에 의해 생성된 표현들은 조인트 네트워크(230)에 의해 결합된다. 그런 다음, 조인트 네트워크는 다음 출력 심볼에 대한 분포인 $P(y_i|x_1, \cdots, x_{t_i}, y_0, \ldots, y_{u_{i-1}})$를 예측한다. 다르게 말하면, 조인트 네트워크(230)는, 각각의 출력 단계(예를 들어, 시간 단계)에서, 가능한 음성 인식 가설들에 대한 확률 분포를 생성한다. 여기서, "가능한 음성 인식 가설들(possible speech recognition hypotheses)"은 지정된 자연어의 심볼/문자를 각각 표현하는 제1 출력 레이블 세트, 및 지정된 자연어의 음소를 각각 표현하는 제2 출력 레이블 세트에 대응한다. 따라서, 조인트 네트워크(230)는 미리 결정된 출력 레이블 세트 각각의 발생 가능성(likelihood of occurrence)을 나타내는 값들의 세트를 출력할 수 있다. 이 값들의 세트는 벡터가 될 수 있으며, 출력 레이블 세트에 대한 확률 분포를 나타낼 수 있다. 일부 경우들에서는, 출력 레이블들이 제1 세트의 자소(grapheme)들(예를 들어, 개별 문자들, 및 잠재적으로 구두점 및 다른 심볼들) 및 제2 세트의 음소들이지만, 출력 레이블 세트는 그와 같이 제한되지 않는다. 조인트 네트워크(230)의 출력 분포는 상이한 출력 레이블들 각각에 대한 사후 확률 값(posterior probability value)을 포함할 수 있다. 따라서, 상이한 자소들 또는 다른 심볼들을 표현하는 100개의 상이한 출력 레이블이 있는 경우, 조인트 네트워크(230)의 출력 y_i는 각각의 출력 레이블에 대해 하나씩 100개의 상이한 확률 값을 포함할 수 있다. 그런 다음, 확률 분포는 전사(116)를 결정하기 위해 (예를 들어, 소프트맥스 계층(240)에 의해) 빔 검색 프로세스에서 후보 정자법 요소(candidate orthgraphic element)들(예를 들어, 자소들, 워드피스들, 워드들, 음소들)을 선택하고 이들에 스코어들을 할당하는 데 사용될 수 있다.

[0030] 소프트맥스 계층(240)은 대응하는 출력 단계에서 모델(200)에 의해 예측된 다음 출력 심볼로서 분포에서 가장 높은 확률을 갖는 출력 레이블/심볼을 선택하기 위해 임의의 기술을 채택할 수 있다. 이러한 방식으로, RNN-T 모델(200)은 조건부 독립 가정을 하지 않고, 오히려, 각각의 심볼의 예측이 음향들뿐만 아니라 지금까지 출력된 레이블들의 시퀀스에 대해서도 조건화된다. RNN-T 모델(200)은 출력 심볼이 미래의 음향 프레임들(110)과 독립적이라고 가정하며, 이는 RNN-T 모델이 스트리밍 방식으로 채택될 수 있게 한다.

[0031] 일부 예들에서, RNN-T 모델(200)의 인코더 네트워크(210)는 8개의 2,048-차원 LSTM 계층으로 구성되고, 각각은 640-차원 투영 계층이 뒤따른다. 모델 레이턴시를 감소시키기 위해 인코더의 제2 LSTM 계층 뒤에 2의 감소 계수(reduction factor)를 갖는 시간-감소 계층이 삽입될 수 있다. 예측 네트워크(220)는 2개의 2,048-차원 LSTM 계층을 가질 수 있으며, 이들 각각은 또한 640-차원 투영 계층이 뒤따른다. 마지막으로, 조인트 네트워크(230)는 또한 640개의 숨겨진 유닛을 가질 수 있고, 이들은 4,096개의 소프트맥스 출력이 뒤따른다. 구체적으로, 출력 유닛들은 41개의 컨텍스트-독립적 음소들을 포함하고, 나머지는 워드피스들이다.

[0032] 다시 도 1을 참조하면, ASR 시스템(100)의 바이어싱 컴포넌트(115)는 외국어 음소들로 바이어싱될 외국어로 된 바이어싱 용어 리스트(105)로부터의 용어들을 토큰화하도록 구성되는 토크나이저(tokenizer)(121), 토큰화된 용어들의 외국어 음소들을 단일 언어, 예를 들어, 미국식 영어와 연관되는 유사한 음소들에 매핑하도록 구성되는 음소 매퍼(123)를 포함한다. 음소 매퍼(123)는 인간이 생성한 소스-언어 대 타겟-언어 음소 쌍들을 포함하는 사전에 의해 표현될 수 있으며, X-SAMPA 음소 세트는 모든 언어들에 사용된다. 특히, 음소 매퍼(123)는 워드피스-음소 모델(200)이 단일 언어, 예를 들어, 미국식 영어와 연관되는 음소들만을 포함할 때 유용하다.

[0033] 예를 들어, 내비게이션 쿼리 "directions to Crétteil"에 대한 발화(106), 및 프랑스어 워드 "Crétteil"이 바이어싱 용어 리스트(105)에 있다는 가정이 주어지면, "Crétteil"은 먼저 토크나이저(121)에 의해 "kRetEj"와 같은 프랑스어 음소들로 토큰화된 후, 음소-레벨 바이어싱 FST(300)를 생성하는 데 사용하

기 위해 음소 매퍼(123)에 의해 영어 음소들 "k r\ E t E j "에 매핑된다. 워드피스-음소 모델(200)은 모델링 유닛들로서 단일 언어, 예를 들어, 미국식 영어로부터의 음소들만을 포함하기 때문에, 음소 매핑이 사용된다.

[0034] 본 개시내용은 바이어싱 용어 리스트(105)에 어떤 용어들이 포함되는지, 또는 바이어싱 용어 리스트(105)에 포함하기 위해 용어들이 선택되는 방법에 제한되지 않는다. 바이어싱 용어 리스트(105)는 관련 컨텍스트에 기초하여 동적으로 업데이트될 수 있다. 예를 들어, 컨텍스트 정보는 어떤 애플리케이션들이 열려 있고 사용자 디바이스(102) 상에서 사용 중인지, 사용자의 연락처 리스트로부터의 연락처 이름들, 사용자(110)의 미디어 라이브러리에 있는 아티스트/앨범 이름들, 사용자(110)의 위치 등을 나타낼 수 있다. 예를 들어, 사용자(102)는 미국식 영어를 말할 수 있고, 내비게이션/맵 애플리케이션이 사용자 디바이스(102) 상에서 열려 있고 사용자(102)의 위치가 프랑스에 있다는 것을 나타내는 컨텍스트 정보에 기초하여, 바이어싱 용어 리스트(105)는 프랑스어로 된 도시 및/또는 지역 이름들과 연관된 용어들을 포함할 수 있다.

[0035] 바이어싱 컴포넌트(115)는 또한 바이어싱 용어 리스트(105)의 외국어(예를 들어, 프랑스어) 용어들 각각을 표현하는 원어(예를 들어, 미국식 영어)의 음소 시퀀스들에 기초하여 바이어싱 FST(300)를 생성하도록 구성되는 음소-레벨 바이어싱 FST 생성기(125)를 포함한다. 일부 예들에서, 바이어싱 생성기(125)는 음소 레벨에서 가중치들을 할당하기 위해 가중치 푸시를 사용하고, 과잉-바이어싱(over-biasing)을 피하기 위해 실패 호(failure arc)들을 추가한다. 일부 구현들에서는, 디코딩에서, 모든 바이어싱 워드들이 동일한 가중치를 갖는 각각의 호를 갖는 컨텍스트 FST를 구성하는 데 사용된다. 이러한 가중치들은 상이한 모델들에 대해 독립적으로 튜닝될 수 있다.

[0036] 음성 인식기(200)는 바이어싱 컴포넌트(115)에 의해 생성된 바이어싱 FST(300)를 사용하여 워드피스-음소 모델(200)에 의해 출력된 음소들을 재스코어링하고, 디코더 그래프(400)는 전사(116)에 포함하기 위한 워드피스들을 생성하기 위해 바이어싱 FST(300)로부터 재스코어링된 음소들 및 워드피스-음소 모델(200)에 의해 출력된 워드피스들을 소비한다. 디코더 그래프(400)는 발화(106)에 대한 하나 이상의 후보 전사를 결정하는 빔 검색 디코딩 프로세스에 대응할 수 있다.

[0037] 일부 예들에서는, 모델(200)에 의한 디코딩 동안, 바이어싱 FST(300)는 워드피스-음소 모델(200)에 의해 출력된 영어 음소 심볼들을 소비할 수 있고, 외국어 어휘집 및 음소 매핑 즉, $\text{"k r\ E t E j"} \rightarrow \text{Créteil}$ 을 사용하여 워드피스들을 생성한다. 디코더 그래프(400)에 의해 출력되는 워드피스들은 연쇄기(134)에 의해 사용자 디바이스(102)의 다른 컴포넌트들, 예를 들어, 사용자 인터페이스 생성기(107) 또는 다른 자연어 프로세싱 컴포넌트들에 출력되는 전사(116)의 워드들로 연쇄된다.

[0038] 도 3은 음소 레벨에서 "Créteil"이라는 워드에 대한 예시적인 바이어싱 FST(300)를 도시한다. 그런 다음, 바이어싱 FST를 사용하여 아래 수학식 (1)을 사용하여 워드피스-음소 모델의 음소 출력들을 온 더 플라이(on the fly)로 리스코어링한다.

수학식 1

[0039]
$$y^* = \arg\max_{y} \log P(y|x) + \lambda \log P_C(y)$$

[0040] 수학식 (1)에서, x는 음향 관찰들을 나타내고, y는 서브워드 유닛 시퀀스를나타내고, P는 E2E 모델로부터의 확률 추정을 나타내고, P_c는 바이어싱 재스코어링 확률이다. λ는 재스코어링에서 컨텍스트 LM의 가중치를 제어한다.

[0041] 다시 도 1을 참조하면, 워드피스-음소 모델(200)은 모델링 유닛들로서 워드피스들뿐만 아니라 음소들을 통합하고, 바이어싱 용어 리스트(105)의 외국어 용어들에 대한 컨텍스트 바이어싱을 위해 바이어싱 FST(300)를 사용한다. 전체-음소 모델과 대조적으로, 음소들 및 워드피스들을 둘 다 모델링하는 워드피스-음소 모델(200)은 일반 워드들을 인식할 때 회귀(regression)들을 완화한다.

[0042] 모델(200)이 스테이지 A에서 트레이닝된 후, 스테이지 B에서, 사용자(110)는 발화(106) "directions to

Crétеil"를 디바이스(102)에 말한다. 스테이지 C에서, 오디오 서브시스템(103)은, 예를 들어, 마이크로폰을 사용하여, 발화를 수신하고, 수신된 발화를 일련의 파라미터화된 입력 음향 프레임들(104)로 변환한다. 예를 들어, 파라미터화된 입력 음향 프레임들(104)은 80-차원 log-Mel 피처들을 각각 포함할 수 있으며, 이들은 짧은, 예를 들어, 25ms의 윈도우로 계산되고, 몇 밀리초마다, 예를 들어, 10밀리초마다 시프트될 수 있다.

[0043] 스테이지 D에서, ASR 시스템(100)은 전술한 바와 같이 파라미터화된 입력 음향 프레임들을 프로세싱하고, 컨텍스트적으로 바이어싱된 전사(116), 즉, 텍스트 "directions to Crétеil"을 출력한다. 스테이지 E에서, 사용자 인터페이스 생성기 시스템(107)은 전사의 표현을 포함하는 그래픽 사용자 인터페이스(136)를 위한 컴퓨터 코드를 생성하고, 스테이지 F에서, 사용자 인터페이스(136) 상에 디스플레이하기 위해 해당 컴퓨터 코드를 모바일 디바이스(102)에 송신한다.

[0044] ASR 시스템(100)에 의해 수행되는 추가 세부사항들은 스테이지 D 동안 발생할 수 있다. 예를 들어, 스테이지 D' 동안, 바이어싱 컴포넌트(115)는 "Crétеil"이라는 용어를 포함하는 바이어싱 용어들의 리스트(120)를 수신하는 것에 기초하여 바이어싱 FST(300)를 생성한다. 스테이지 D''에서, 음성 인식기(150)의 트레이닝된 워드피스-음소 모델(200)은 사용자(110)의 발화(106)에 기초하여 워드피스들 및 대응하는 음소 시퀀스들 둘 다에 대한 음성 인식 스코어들을 생성하며, 음소들에 대한 음성 인식 스코어들은 바이어싱 FST(300)에 의해 재스코어링 및 재매핑되고, 스테이지 D'''에서, 전사(116)의 출력을 위한 워드피스들을 생성하기 위해 워드피스들 및 재스코어링된/재매핑된 음소들이 디코더 그래프(400)에 의해 소비된다. 디코더 그래프(400) 및 연쇄기(134)는 컨텍스트적으로 바이어싱된 전사(116)를 생성하고, 출력을 위한 전사를, 예를 들어, 사용자 디바이스(102)의 GUI(136)에 디스플레이하기 위해 사용자 인터페이스 생성기 시스템(107)에 제공한다. 특히, 디코더 그래프(400)는 바이어싱 FST(300)가 바이어싱 용어 리스트(105)의 용어들 중 임의의 것에 대응하는 음소 시퀀스들을 재스코어링한 후에 실행된다. 이와 같이, 바이어싱 용어 리스트(105)의 외국어 워드들에 대응하는 낮은 음성 인식 스코어들을 갖는 워드피스들은 조기에 가지치기되지 않는다(not pruned).

[0045] 테스트 동안, 음성 인식기(150)가 워드피스-음소 모델(200) 및 바이어싱 FST(300)를 채택하여 바이어싱 용어 리스트(105)의 용어들에 대한 인식 결과들을 컨텍스트적으로 바이어싱하면 자소-전용 바이어싱 모델들 및 워드피스-전용 바이어싱 모델들 둘 다보다 현저하게 우수한 WER 레이트로 외국어 워드들을 성공적으로 인식하는 것으로 나타났다. 워드피스-음소 모델(200)은 또한 모델 확장성 이슈들 없이 바이어싱을 위해 다른 외국어들에 직접 적용될 수 있다는 장점을 갖는다.

[0046] 도 4는 음성 인식기(150)가 음성 인식 결과들을 컨텍스트적으로 바이어싱하기 위해 실행하는 예시적인 디코딩 그래프(400)를 도시한다. 구체적으로, 예시적인 디코딩 그래프(400)는 영어 교차 언어 발음 "k r\ E S"을 갖는 워드 "crèche"(영어로, "daycare") 및 발음 "k r\ E t E j"을 갖는 워드 "crétеil"(프랑스의 도시)에 대한 디코딩을 도시한다. 명확성을 위해, 상태 0에 대한 대부분의 워드피스들은 생략된다.

[0047] 디코딩 그래프(400)는 워드피스-음소 모델(200)로부터 출력된 음소들 및 워드피스들 둘 다를 입력으로서 수신하도록 구성된다. 음성 디코딩 프로세스는 출력들로서 워드들을 생성하기 위해 디코딩 그래프(400)를 검색한다. 도시된 예에서는, 디코딩 FST가 상태 0 주변에 워드피스 루프들을 가지고 있지만, 발음 FST, 즉, 입력들로서 음소들을 갖고 출력들로서 대응하는 워드피스들을 갖는 프리픽스 트리(prefix tree)를 포함하는 상태 1 내지 14를 갖는다. 발음 FST는 모든 바이어싱 용어들에 대한 바이어싱에 사용된 것과 동일한 발음들을 사용하여 구성된다. 항상 워드피스들인 최종 출력 심볼들은 (예를 들어, 도 1의 연쇄기(134)에 의해) 워드들로 연쇄된다.

[0048] 도 4의 디코딩 그래프(400)는 전체 디코딩 전략에 대한 두 가지 추가 향상 사항들을 제공한다. 첫째, 디코딩 그래프(400)의 특성을 감안할 때, 동일한 비용으로 동일한 입력들을 소비하지만 동일한 출력들을 갖지 않는다는 여러 가설들이 있을 수 있다. 예를 들어, 상태 7에서 끝나는 가설은 상태 9에서 끝나는 가설과 동일한 비용을 가질 것이다. 이는 빔이 모두 동일한 많은 가설들에 의해 채워지기 때문에 이슈들을 일으킨다. 따라서, 본 명세서에서 설명되는 강화된 ASR 기술들은 상태 9에서 끝나는 가설만을 유지함으로써 빔에 대해 가지치기한다.

[0049] 두 번째 개선 사항은 병합된 경로들에 관한 것이다. 트레이닝 및 디코딩의 특성을 감안할 때, 주어진 워드가 워드피스들로 직접 출력될 수도 있고, 또는 음소들로부터 워드피스들로 변환될 수도 있다. 등가의 가설들은 그들의 확률들을 추가하고 총 확률을 가장 가능성이 높은 가설에 할당하고 나머지들은 빔으로부터 드롭함으로써

추적 및 재결합된다.

[0050] 워드피스-음소 모델(200)의 바이어싱 결과들을 자소-전용 모델 및 워드피스-전용 모델과 비교하기 위해 테스트가 수행되었다. 후자의 두 모델은 워드피스-음소 모델(200)과 동일한 구조를 가지는데, 차이는 자소 모델은 76개의 자소를 출력들로서 갖고, 워드피스 모델은 4,096개의 워드피스를 갖는다는 점이다. 이 차이는 자소 모델 및 워드피스 모델에 대해 각각 약 117M개 및 120M개의 파라미터로 이어진다. 두 모델의 출력 심볼들은 영어로 되어 있으며, 이들은 모든 영어 데이터를 사용하여 트레이닝된다는 점에 유의하도록 한다. 이 두 모델의 경우, 프랑스어 바이어싱 워드들의 영어 음역된 버전들을 사용하여 자소 레벨 또는 워드피스 레벨에서만 바이어싱이 수행된다.

[0051] 일반적으로, 테스트는 세 가지 모델 모두가 바이어싱 없이도 유사하게 수행함을 나타낸다. 이는 장소 이름들이 프랑스어로 되어 있고 이들이 트레이닝에서 보여진 적이 없기 때문이며, 즉, 워드 OOV 레이트가 거의 100%이다. 또한, 모든 모델들은 바이어싱에 의해 상당히 더 우수하게 수행된다. 바이어싱이 없는 사례와 비교할 때, WER 감소들은 중요하다.

[0052] 상이한 바이어싱 전략들을 비교하면, 워드피스-음소 모델(200)은 자소 모델 및 워드피스 모델 둘 다보다 상당히 더 우수하게 동작하는 최상의 성능을 나타내었다. 워드피스-음소 모델의 우수한 성능은 OOV 워드들에 대한 음소들의 견고성에 기인한다. 워드피스-음소 모델(200)은 모델링 유닛들로서 워드피스들 및 음소들 둘 다를 포함하기 때문에, 음소 FST 외에 워드피스 FST를 구축함으로써 음소-기반 바이어싱 외에 워드피스 바이어싱이 수행될 수 있다. 이 워드피스 FST의 추가는 WER을 추가로 감소시키는 것으로 입증되었으며, 이는 해당 워드피스 바이어싱과 음소 바이어싱이 서로 보완적임을 나타낸다. 음소 및 워드피스 바이어싱 둘 다에 사용되는 가중치들은 동일할 수도 있고 상이할 수 있다. 관찰들은 더 긴 유닛들을 매칭시킬 때 희소성(sparsity) 이슈들이 있기 때문에 워드피스들이 자소들보다 더 우수하게 수행될 수 있다고 나타내었다.

[0053] 테스트는 바이어싱이 외국 장소 이름들을 인식하는 데 도움이 될 수 있다고 나타내었다. 예를 들어, 바이어싱은 올바른 프랑스어 워드들을 생성하는 데 도움이 되며, 반대로, 바이어싱이 없을 때에는 음성학적으로 유사하지만 잘못된 영어 워드들이 생성된다. 오류들은 종종 프랑스어에서 음성학적으로 유사한 워드들로 인해 발생한다.

[0054] 바이어싱 없는 시나리오들에서 회귀가 없음을 더 잘 보장하기 위해, 정규 영어 발화들을 디코딩하기 위한 세 가지 모델이 비교되었다. 디코딩에서, 비어있는 바이어싱 구문들의 리스트를 사용함으로써 바이어싱 메커니즘이 해제되었다. 테스트는 워드피스 모델이 자소 모델보다 더 우수하게 수행하는 것으로 나타내었다. 워드피스-음소 모델은 자소 모델보다 약간 더 우수하게 수행하는데, 이는 트레이닝 동안 워드피스들의 빈도가 더 높기 때문일 수 있다. 워드피스 모델과 비교하여, 워드피스-음소 모델은 매우 약간의 열화가 있다. 이것은 모델링에 전화들이 도입되었기 때문이다. 회귀를 향상시키기 위한 한 가지 잠재적인 접근 방식은 워드피스-기반 재스코어링과 유사하게 음소들에 대한 영어 외부 언어 모델을 재스코어링에 통합하는 것이다. 그러나, 회귀는 모든 음소 모델보다 훨씬 더 작다.

[0055] 도 5는 바이어싱 용어 리스트의 외국어 용어들에 대한 전사를 컨텍스트적으로 바이어싱하는 방법에 대한 예시적인 동작들의 배열의 흐름도이다. 동작(502)에서, 방법(500)은 제1 언어의 원어민(110)에 의해 말해진 발화(106)를 인코딩하는 오디오 데이터를 수신하는 동작을 포함한다. 발화(106)는 제1 언어와 상이한 제2 언어의 하나 이상의 외국어 워드를 포함할 수 있다. 동작(504)에서, 방법(500)은 제2 언어로 된 하나 이상의 용어를 포함하는 바이어싱 용어 리스트(105)를 수신하는 동작을 포함한다.

[0056] 동작(506)에서, 방법(500)은 또한 음성 인식 모델(200)을 사용하여 제1 언어의 워드피스들 및 대응하는 음소 시퀀스들 둘 다에 대한 음성 인식 스코어들을 생성하기 위해 오디오 데이터로부터 도출된 음향 피처들(104)을 프로세싱하는 동작을 포함한다. 동작(508)에서, 방법(500)은 또한 바이어싱 용어 리스트의 하나 이상의 용어에 기초하여 음소 시퀀스들에 대한 음성 인식 스코어들을 재스코어링하는 동작을 포함한다. 동작(506)에서, 방법(500)은, 워드피스들에 대한 음성 인식 스코어들 및 음소 시퀀스들에 대한 재스코어링된 음성 인식 스코어들을 사용하여, 발화(106)에 대한 전사(116)를 생성하기 위해 디코딩 그래프(400)를 실행하는 동작을 포함한다.

[0057] 소프트웨어 애플리케이션(즉, 소프트웨어 리소스)은 컴퓨팅 디바이스로 하여금 태스크를 수행하게 하는 컴퓨터 소프트웨어를 지칭할 수 있다. 일부 예들에서, 소프트웨어 애플리케이션은 "애플리케이션", "앱" 또는 "프로그램"으로 지칭될 수 있다. 예시적인 애플리케이션들은 시스템 진단 애플리케이션들, 시스템 관리 애플리케이션들, 시스템 유지보수 애플리케이션들, 워드 프로세싱 애플리케이션들, 스프레드시트 애플리케이션들, 메시징 애

플리케이션들, 미디어 스트리밍 애플리케이션들, 소셜 네트워킹 애플리케이션들 및 게임 애플리케이션들을 포함하지만, 이에 제한되지 않는다.

[0058] 비-일시적 메모리는 컴퓨팅 디바이스에 의한 사용을 위해 일시적으로 또는 영구적으로 프로그램들(예를 들어, 명령어들의 시퀀스들) 또는 데이터(예를 들어, 프로그램 상태 정보)을 저장하는 데 사용되는 물리적 디바이스들일 수 있다. 비-일시적 메모리는 휘발성 및/또는 비-휘발성 어드레스 지정 가능 반도체 메모리일 수 있다. 비-휘발성 메모리의 예들은 플래시 메모리 및 판독 전용 메모리(read-only memory)(ROM)/프로그래밍 가능 판독 전용 메모리(programmable read-only memory)(PROM)/소거 가능한 프로그래밍 가능 판독 전용 메모리(erasable programmable read-only memory)(EPROM)/전자적으로 소거 가능한 프로그래밍 가능 판독 전용 메모리(electronically erasable programmable read-only memory)(EEPROM)(예를 들어, 통상적으로 부팅 프로그램들과 같은 펌웨어에 사용됨)를 포함하지만, 이에 제한되지 않는다. 휘발성 메모리의 예들은 랜덤 액세스 메모리(random access memory)(RAM), 동적 랜덤 액세스 메모리(dynamic random access memory)(DRAM), 정적 랜덤 액세스 메모리(static random access memory)(SRAM), 상 변화 메모리(phase change memory)(PCM)뿐만 아니라, 디스크들 또는 테이프들을 포함하지만, 이에 제한되지 않는다.

[0059] 도 6은 본 문헌에서 설명되는 시스템들 및 방법들을 구현하는 데 사용될 수 있는 예시적인 컴퓨팅 디바이스(600)의 개략도이다. 컴퓨팅 디바이스(600)는 랩탑들, 데스크탑들, 워크스테이션들, 퍼스널 디지털 어시스턴트들, 서버들, 블레이드 서버들, 메인프레임들 및 기타 적절한 컴퓨터들과 같은 다양한 형태들의 디지털 컴퓨터들을 나타내도록 의도된다. 본 명세서에서 도시된 컴포넌트들, 이들의 연결들 및 관계들, 및 이들의 기능들은 예시에 불과한 것임을 의미하며, 본 문헌에서 설명 및/또는 청구되는 발명들의 구현들을 제한하는 것으로 의미하지 않는다.

[0060] 컴퓨팅 디바이스(600)는 프로세서(610), 메모리(620), 스토리지 디바이스(630), 메모리(620) 및 고속 확장 포트들(650)에 연결되는 고속 인터페이스/컨트롤러(640), 및 저속 버스(670) 및 스토리지 디바이스(630)에 연결되는 저속 인터페이스/컨트롤러(660)를 포함한다. 컴포넌트들(610, 620, 630, 640, 650 및 660) 각각은 다양한 버스들을 사용하여 상호 연결되며, 공통 마더보드 상에 또는 적절한 다른 방식들로 마운팅될 수 있다. 프로세서(610)는 고속 인터페이스(640)에 커플링되는 디스플레이(680)와 같은 외부 입/출력 디바이스 상에 그래픽 사용자 인터페이스(graphical user interface)(GUI)에 대한 그래픽 정보를 디스플레이하기 위해 메모리(620)에 또는 스토리지 디바이스(630) 상에 저장되는 명령어들을 포함하여 컴퓨팅 디바이스(600) 내에서 실행하기 위한 명령어들을 프로세싱할 수 있다. 다른 구현들에서, 다수의 프로세서들 및/또는 다수의 버스들은 적절한 경우에 다수의 메모리들 및 다수의 타입들의 메모리와 함께 사용될 수 있다. 또한, 다수의 컴퓨팅 디바이스들(600)이 연결될 수 있으며, 각각의 디바이스는 (예를 들어, 서버 뱅크, 블레이드 서버들의 그룹, 또는 멀티-프로세서 시스템으로서) 필요한 동작들의 일부들을 제공한다.

[0061] 메모리(620)는 정보를 컴퓨팅 디바이스(600) 내에 비-일시적으로 저장한다. 메모리(620)는 컴퓨터 판독 가능 매체, 휘발성 메모리 유닛(들), 또는 비-휘발성 메모리 유닛(들)일 수 있다. 비-일시적 메모리(620)는 컴퓨팅 디바이스(600)에 의한 사용을 위해 프로그램들(예를 들어, 명령어들의 시퀀스들) 또는 데이터(예를 들어, 프로그램 상태 정보)를 일시적으로 또는 영구적으로 저장하는 데 사용되는 물리적 디바이스들일 수 있다. 비-휘발성 메모리의 예들은 플래시 메모리 및 판독 전용 메모리(ROM)/프로그래밍 가능 판독 전용 메모리(PROM)/소거 가능한 프로그래밍 가능 판독 전용 메모리(EPROM)/전자적으로 소거 가능한 프로그래밍 가능 판독 전용 메모리(EEPROM)(예를 들어, 통상적으로 부팅 프로그램들과 같은 펌웨어에 사용됨)를 포함하지만, 이에 제한되지 않는다. 휘발성 메모리의 예들은 랜덤 액세스 메모리(RAM), 동적 랜덤 액세스 메모리(DRAM), 정적 랜덤 액세스 메모리(SRAM), 상 변화 메모리(PCM)뿐만 아니라, 디스크들 또는 테이프들을 포함하지만, 이에 제한되지 않는다.

[0062] 스토리지 디바이스(630)는 컴퓨팅 디바이스(600)를 위한 대용량 스토리지를 제공할 수 있다. 일부 구현들에서, 스토리지 디바이스(630)는 컴퓨터 판독 가능 매체이다. 다양한 상이한 구현들에서, 스토리지 디바이스(630)는 플로피 디스크 디바이스, 하드 디스크 디바이스, 광 디스크 디바이스, 또는 테이프 디바이스, 플래시 메모리 또는 다른 유사한 솔리드 상태 메모리 디바이스, 또는 스토리지 영역 네트워크 또는 기타 구성들의 디바이스들을 포함한 디바이스들의 어레이일 수 있다. 추가 구현들에서, 컴퓨터 프로그램 제품은 정보 캐리어에 유형적으로(tangibly) 구체화된다. 컴퓨터 프로그램 제품은, 실행될 때, 위에서 설명된 것들과 같은 하나 이상의 방법을 수행하는 명령어들을 포함한다. 정보 캐리어는 메모리(620), 스토리지 디바이스(630), 또는 프로세서(610) 상의 메모리와 같은 컴퓨터 또는 머신 판독 가능 매체이다.

[0063] 고속 컨트롤러(640)는 컴퓨팅 디바이스(600)에 대한 대역폭-집약적 동작들을 관리하는 반면, 저속 컨트롤러

(660)는 더 낮은 대역폭-집약적 동작들을 관리한다. 이러한 직무들의 할당은 예시일 뿐이다. 일부 구현들에서, 고속 컨트롤러(640)는 메모리(620), 디스플레이(680)에(예를 들어, 그래픽 프로세서 또는 가속기를 통해), 및 다양한 확장 카드들(도시 생략)을 수용할 수 있는 고속 확장 포트들(650)에 커플링된다. 일부 구현들에서, 저속 컨트롤러(660)는 스토리지 디바이스(630) 및 저속 확장 포트(690)에 커플링된다. 다양한 통신 포트들(예를 들어, USB, 블루투스, 이더넷, 무선 이더넷)을 포함할 수 있는 저속 확장 포트(690)는 키보드, 포인팅 디바이스, 스캐너와 같은 하나 이상의 입/출력 디바이스, 또는, 예를 들어, 네트워크 어댑터를 통해 스위치 또는 라우터와 같은 네트워킹 디바이스에 커플링될 수 있다.

[0064] 컴퓨팅 디바이스(600)는 도면에 도시된 바와 같이 다수의 상이한 형태들로 구현될 수 있다. 예를 들어, 이것은 표준 서버(600a)로서 또는 이러한 서버들(600a)의 그룹에서 여러 번, 랩톱 컴퓨터(600b)로서, 또는 랙 서버 시스템(600c)의 일부로서 구현될 수 있다.

[0065] 본 명세서에서 설명되는 시스템들 및 기술들의 다양한 구현들은 디지털 전자 및/또는 광 회로망, 집적 회로망, 특별히 설계된 ASIC(application specific integrated circuit)들, 컴퓨터 하드웨어, 펌웨어, 소프트웨어, 및/또는 이들의 조합으로 실현될 수 있다. 이러한 다양한 구현들은 스토리지 시스템, 적어도 하나의 입력 디바이스 및 적어도 하나의 출력 디바이스로부터 데이터 및 명령어들을 수신하고 이에 데이터 및 명령어들을 송신하기 위해 커플링되는 특수 목적 또는 범용일 수 있는 적어도 하나의 프로그래밍 가능 프로세서를 포함하는 프로그래밍 가능 시스템 상에서 실행 가능하고/하거나 해석 가능한 하나 이상의 컴퓨터 프로그램에서의 구현을 포함할 수 있다.

[0066] 이러한 컴퓨터 프로그램들(프로그램들, 소프트웨어, 소프트웨어 애플리케이션들 또는 코드로도 알려짐)은 프로그래밍 가능 프로세서를 위한 머신 명령어들을 포함하고, 고-레벨 절차 및/또는 객체-지향 프로그래밍 언어로 및/또는 어셈블리/머신 언어로 구현될 수 있다. 본 명세서에서 사용되는 바와 같이, "머신 판독 가능 매체" 및 "컴퓨터 판독 가능 매체"라는 용어들은 머신 판독 가능 신호로서 머신 명령어들을 수신하는 머신 판독 가능 매체를 포함하여 프로그래밍 가능 프로세서에 머신 명령어들 및/또는 데이터를 제공하는 데 사용되는 임의의 컴퓨터 프로그램 제품, 비-일시적 컴퓨터 판독 가능 매체, 장치 및/또는 디바이스(예를 들어, 자기 디스크들, 광 디스크들, 메모리, 프로그래밍 가능 로직 디바이스(Programmable Logic Device)들(PLD들)을 지칭한다. "머신 판독 가능 신호"라는 용어는 프로그래밍 가능 프로세서에 머신 명령어들 및/또는 데이터를 제공하는 데 사용되는 임의의 신호를 지칭한다.

[0067] 본 명세서에서 설명되는 프로세스들 및 로직 흐름들은 데이터 프로세싱 하드웨어라고도 지칭되는 하나 이상의 프로그래밍 가능 프로세서에 의해 수행될 수 있으며, 이는 입력 데이터에 대해 동작하고 출력을 생성함으로써 기능들을 수행하기 위해 하나 이상의 컴퓨터 프로그램을 실행한다. 프로세스들 및 로직 흐름들은 예를 들어, FPGA(field programmable gate array) 또는 ASIC(application specific integrated circuit)와 같은 특수 목적 로직 회로망에 의해 수행될 수도 있다. 컴퓨터 프로그램의 실행에 적절한 프로세서들은, 예를 들어, 범용 마이크로프로세서 및 특수 목적 마이크로프로세서 둘 다, 및 임의의 종류의 디지털 컴퓨터의 임의의 하나 이상의 프로세서를 포함한다. 일반적으로, 프로세서는 판독 전용 메모리 또는 랜덤 액세스 메모리 또는 둘 다로부터 명령어들 및 데이터를 수신할 것이다. 컴퓨터의 필수 요소들은 명령어들을 수행하기 위한 프로세서 및 명령어들 및 데이터를 저장하기 위한 하나 이상의 메모리 디바이스이다. 일반적으로, 컴퓨터는 또한 데이터를 저장하기 위한 하나 이상의 대용량 스토리지 디바이스, 예를 들어, 자기, 광자기 디스크(magneto optical disk)들, 또는 광 디스크들을 포함하거나, 또는 이로부터 데이터를 수신하거나 이에 데이터를 전송하거나 또는 둘 다를 수행하기 위해 이에 동작 가능하게 커플링될 것이다. 그러나, 컴퓨터는 이러한 디바이스들을 가질 필요는 없다. 컴퓨터 프로그램 명령어들 및 데이터를 저장하기에 적절한 컴퓨터 판독 가능 매체는, 예를 들어, 반도체 메모리 디바이스들, 예를 들어, EPROM, EEPROM 및 플래시 메모리 디바이스들; 자기 디스크들, 예를 들어, 내부 하드 디스크들 또는 이동식 디스크들; 광자기 디스크들; 및 CD ROM 및 DVD-ROM 디스크들을 포함하여 모든 형태들의 비-휘발성 메모리, 매체 및 메모리 디바이스들을 포함한다. 프로세서 및 메모리는 특수 목적 로직 회로망에 의해 보완되거나 또는 이에 통합될 수 있다.

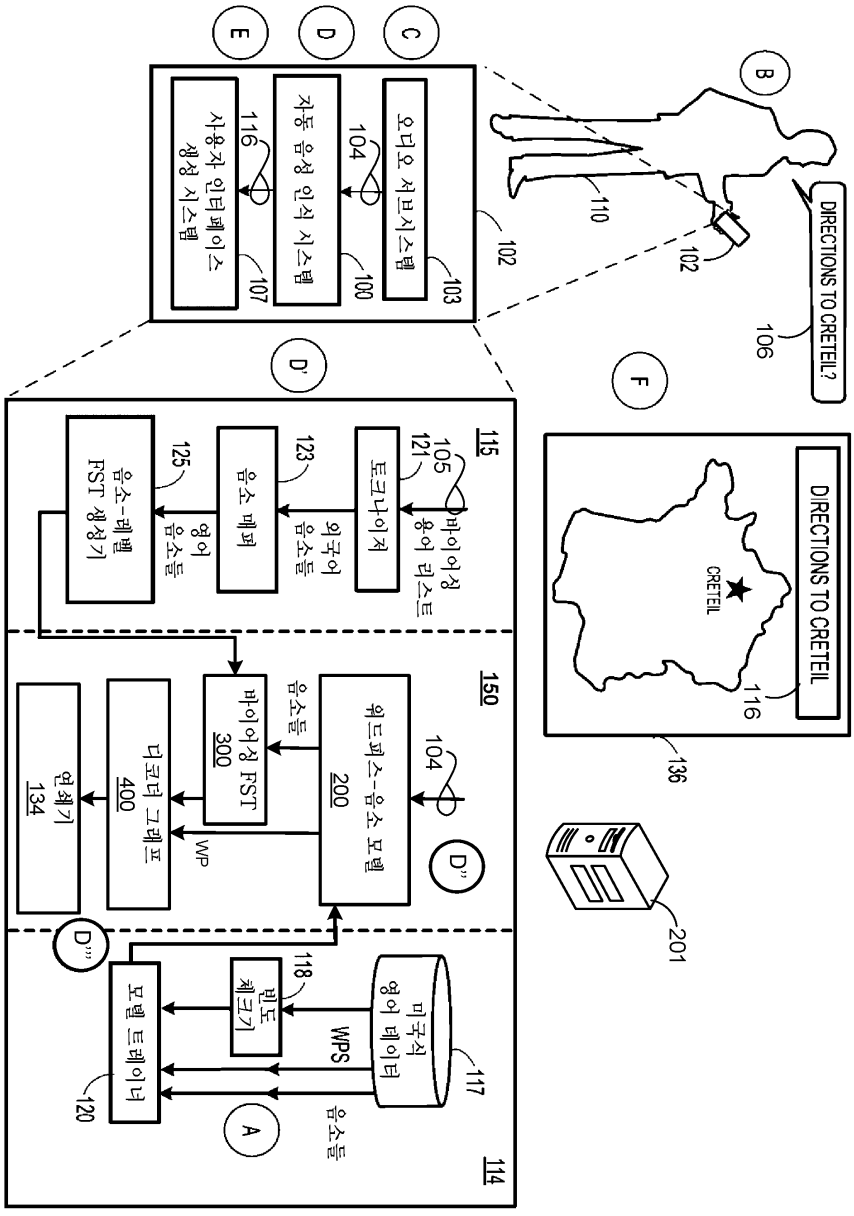
[0068] 사용자와의 상호 작용을 제공하기 위해, 본 개시내용의 하나 이상의 양태는 디스플레이 디바이스, 예를 들어, CRT(cathode ray tube), LCD(liquid crystal display) 모니터, 또는 사용자에게 정보를 디스플레이하기 위한 터치 스크린, 및 임의적으로는 사용자가 컴퓨터에 입력을 제공할 수 있는 키보드 및 포인팅 디바이스, 예를 들어, 마우스 또는 트랙볼을 갖는 컴퓨터 상에서 구현될 수 있다. 다른 종류들의 디바이스들도 사용자와의 상호 작용을 제공하는 데 사용될 수 있으며, 예를 들어, 사용자에게 제공되는 피드백은 임의의 형태의 감각적 피드백, 예를 들어, 시각적 피드백, 청각적 피드백 또는 촉각적 피드백일 수 있고, 사용자로부터의 입력은

음향, 음성 또는 촉각 입력을 포함한 임의의 형태로 수신될 수 있다. 또한, 컴퓨터는 사용자에 의해 사용되는 디바이스로 문서들을 전송하고 이로부터 문서들을 수신함으로써, 예를 들어, 웹 브라우저로부터 수신된 요청들에 응답하여 사용자의 클라이언트 디바이스 상의 웹 브라우저에 웹 페이지들을 전송함으로써 사용자와 상호 작용할 수 있다.

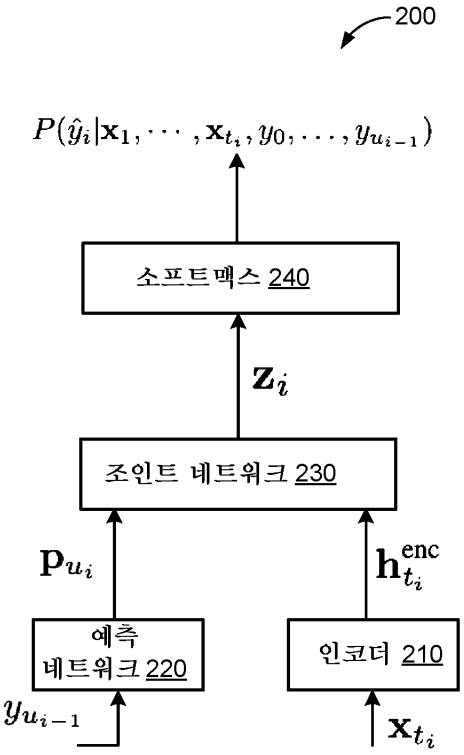
[0069] 다수의 구현들이 설명되었다. 그럼에도 불구하고, 본 개시내용의 사상 및 범주를 벗어나지 않고 다양한 수정들이 이루어질 수 있음이 이해될 것이다. 따라서, 다른 구현들은 다음 청구항의 범위 내에 있다.

도면

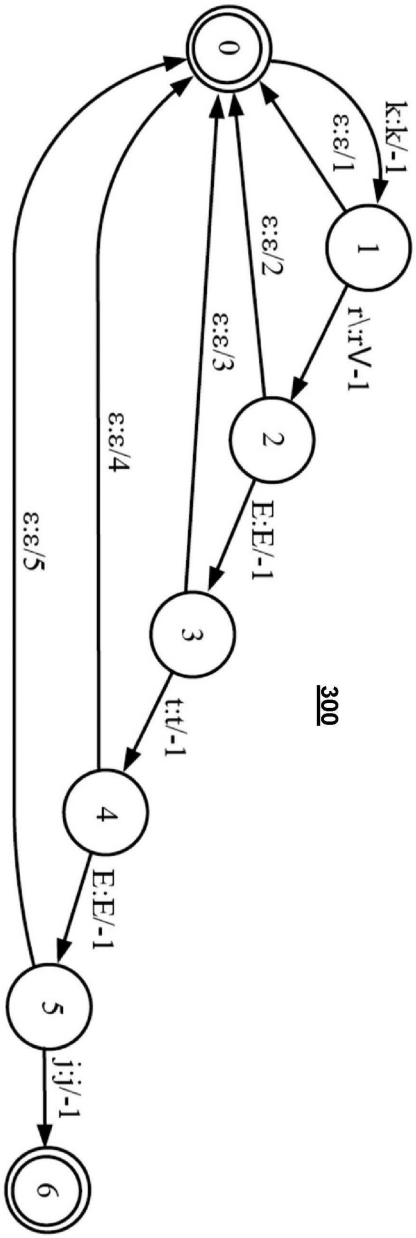
도면1



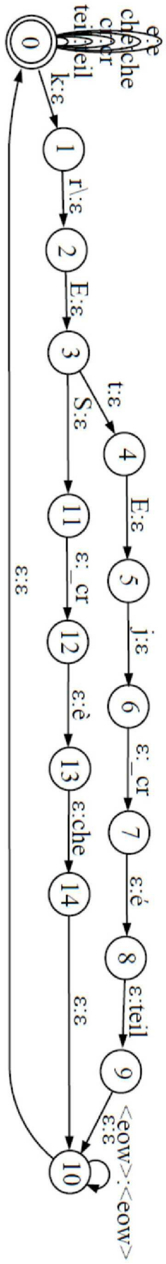
도면2


200
$P(\hat{y}_i|\mathbf{x}_1,\cdots,\mathbf{x}_{t_i},y_0,\ldots,y_{u_{i-1}})$
소프트맥스 240
$\mathbf{z}_i$
조인트 네트워크 230
$\mathbf{p}_{u_i}$
$\mathbf{h}_{t_i}^{\text{enc}}$
예측 네트워크 220
인코더 210
$y_{u_{i-1}}$
$\mathbf{x}_{t_i}$

도면3

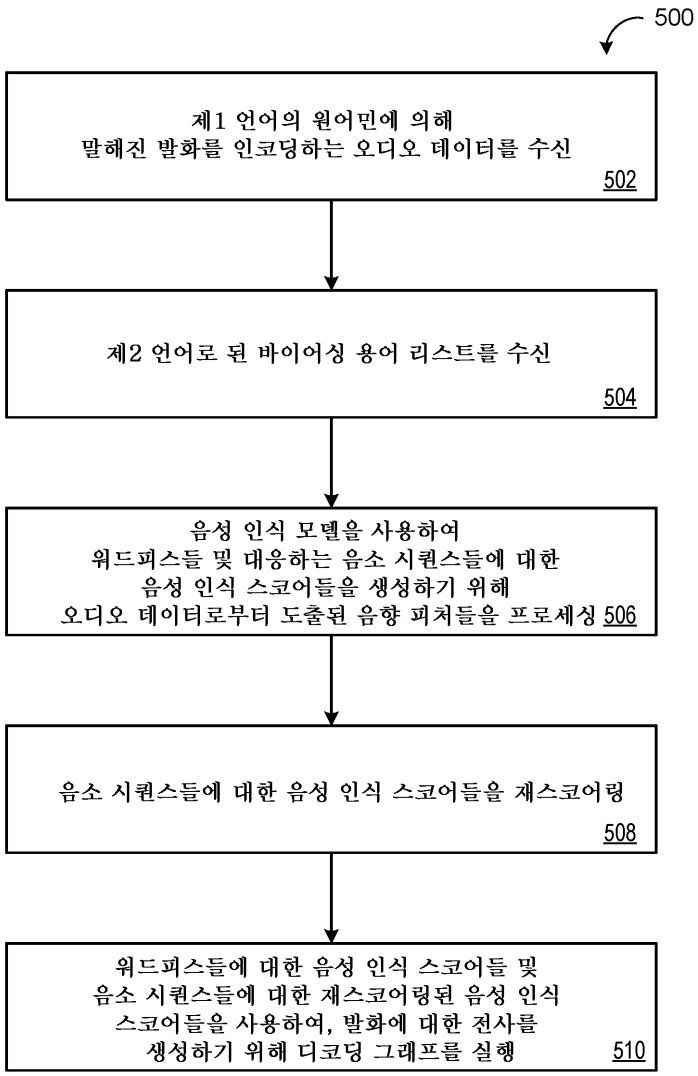


도면4

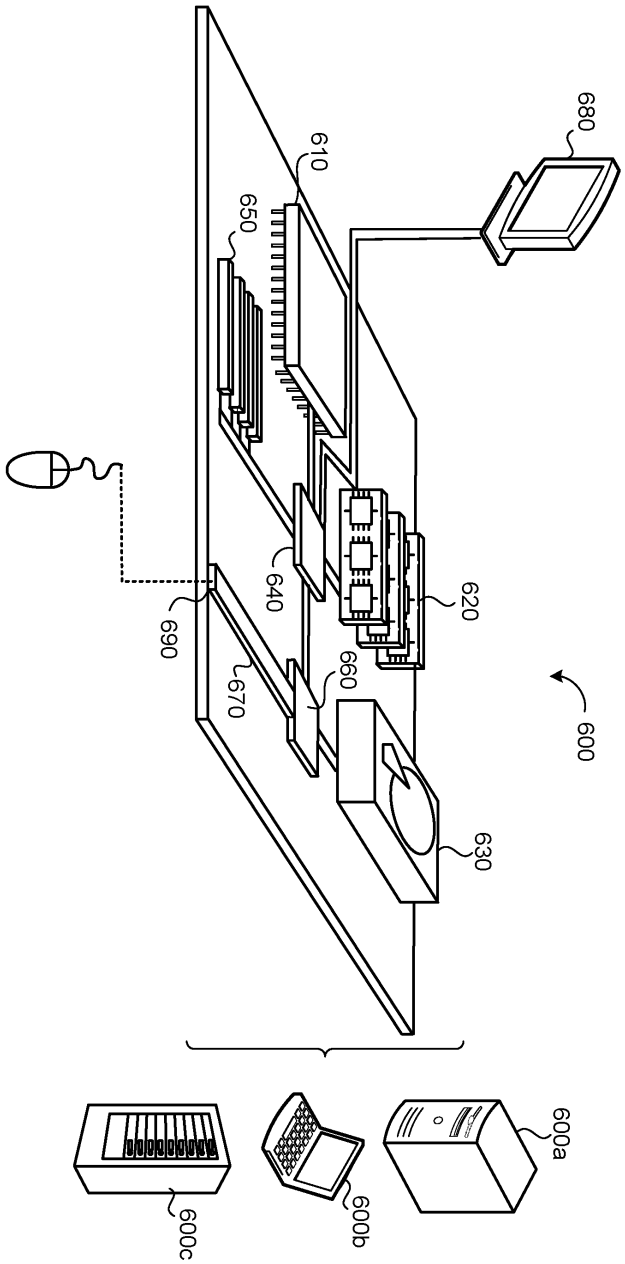

0
k:ε
1
r\:ε
2
E:ε
3
t:ε
S:ε
4
E:ε
5
j:ε
6
ε:_cr
7
ε:é
8
ε:teil
9
11
ε:_cr
12
ε:è
13
ε:che
14
ε:ε
10
<eow>:<eow>

400

도면5


500
제1 언어의 원어민에 의해
말해진 발화를 인코딩하는 오디오 데이터를 수신
502
제2 언어로 된 바이어싱 용어 리스트를 수신
504
음성 인식 모델을 사용하여
워드피스들 및 대응하는 음소 시퀀스들에 대한
음성 인식 스코어들을 생성하기 위해
오디오 데이터로부터 도출된 음향 피처들을 프로세싱 506
음소 시퀀스들에 대한 음성 인식 스코어들을 재스코어링
508
워드피스들에 대한 음성 인식 스코어들 및
음소 시퀀스들에 대한 재스코어링된 음성 인식
스코어들을 사용하여, 발화에 대한 전사를
생성하기 위해 디코딩 그래프를 실행
510

도면6



【심사관 직권보정사항】

【직권보정 1】

【보정항목】 청구범위

【보정세부항목】 청구항 3

【변경전】

제2항에 있어서,

상기 데이터 프로세싱 하드웨어(610)에 의해, 상기 바이어싱 용어 리스트(105)의 각각의 용어를 상기 제2 언어의 대응하는 음소 시퀀스로 토큰화하는 단계;

상기 데이터 프로세싱 하드웨어(610)에 의해, 상기 제2 언어의 각각의 대응하는 음소 시퀀스를 상기 제1 언어의 대응하는 음소 시퀀스에 매핑하는 단계; 및

상기 데이터 프로세싱 하드웨어(610)에 의해, 상기 제1 언어의 각각의 대응하는 음소 시퀀스에 기초하여 상기

바이어싱 FST(300)를 생성하는 단계

를 추가로 포함하는, 방법(500).

【변경후】

제2항에 있어서,

상기 데이터 프로세싱 하드웨어(610)에 의해, 상기 바이어싱 용어 리스트(105)의 각각의 용어를 상기 제2 언어의 대응하는 음소 시퀀스로 토큰화하는 단계;

상기 데이터 프로세싱 하드웨어(610)에 의해, 상기 제2 언어의 각각의 대응하는 음소 시퀀스를 상기 제1 언어의 대응하는 음소 시퀀스에 매핑하는 단계; 및

상기 데이터 프로세싱 하드웨어(610)에 의해, 상기 제1 언어의 각각의 대응하는 음소 시퀀스에 기초하여 상기 바이어싱 FST를 생성하는 단계

를 추가로 포함하는, 방법(500).

【직권보정 2】

【보정항목】 청구범위

【보정세부항목】 청구항 13

【변경전】

제12항에 있어서, 상기 동작들은,

상기 바이어싱 용어 리스트(105)의 각각의 용어를 상기 제2 언어의 대응하는 음소 시퀀스로 토큰화하는 동작;

상기 제2 언어의 각각의 대응하는 음소 시퀀스를 상기 제1 언어의 대응하는 음소 시퀀스에 매핑하는 동작; 및

상기 제1 언어의 각각의 대응하는 음소 시퀀스에 기초하여 상기 바이어싱 FST(300)를 생성하는 동작

을 추가로 포함하는, 시스템(100).

【변경후】

제12항에 있어서, 상기 동작들은,

상기 바이어싱 용어 리스트(105)의 각각의 용어를 상기 제2 언어의 대응하는 음소 시퀀스로 토큰화하는 동작;

상기 제2 언어의 각각의 대응하는 음소 시퀀스를 상기 제1 언어의 대응하는 음소 시퀀스에 매핑하는 동작; 및

상기 제1 언어의 각각의 대응하는 음소 시퀀스에 기초하여 상기 바이어싱 FST를 생성하는 동작

을 추가로 포함하는, 시스템(100).